

CHAP 4: STATISTIQUE BAYESIENNE

Ces notes de cours sont inspirées de plusieurs supports, notamment des notes de M. Rebak.

Ref. biblio : C.P. Robert - The Bayesian Choice : A Decision-theoretic Motivation - Springer.

I. Introduction

→ On se place dans le cadre d'un modèle statistique paramétrique $\{X, \mathcal{A}, \{P_\theta : \theta \in \Theta\}\}$.
La statistique fréquentiste suppose l'existence d'un "vrai" paramètre θ_0 qui génère les données, i.e. $(X_1, \dots, X_n) \sim g(\cdot; \theta_0)$ (g est la densité de probabilité).

Puis on cherche un estimateur $\hat{\theta}_n$ de θ_0 qui a de bonnes propriétés (comme EMV) :

$$\forall n \quad (\hat{\theta}_n - \theta_0) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

L'estimateur $\hat{\theta}_n$ nous donne une estimation ponctuelle pour θ_0 , que l'on pourra compléter par une estimation ensembliste (IC) en utilisant ces résultats asymptotiques. Par exemple, $IC_n(\theta) = \left[\hat{\theta}_n - q_{1-\frac{\alpha}{2}}^{(0,1)} \hat{\sigma}_n ; \hat{\theta}_n + q_{1-\frac{\alpha}{2}}^{(0,1)} \hat{\sigma}_n \right]$ avec $\hat{\sigma}_n$ un estimateur de l'écart-type de $\hat{\theta}_n$.

En résumé, $\hat{\theta}_n$ est une v.a. pour laquelle on ne connaît sur que la loi asymptotique.

→ Dans le cadre bayésien, on ne considère plus le paramètre θ comme déterministe et inconnu. On considère que θ est lui-même une v.a.

II - Fondamentaux de la statistique bayésienne.

En bayésien, on modélise donc également l'incertitude que l'on a sur le paramètre par une loi sur ce paramètre. On se sert de l'expérience accumulée (via les observations x_1, \dots, x_n réalisées) pour mettre à jour cette incertitude.

① - Loi a priori

C'est la loi choisie pour modéliser l'incertitude sur le paramètre. Cette loi a priori permet d'intégrer par exemple en amont d'expert, d'encoder nos connaissances a priori et nos ignorances sur θ . AVANT d'observer des données.

On la note en général π , donc $\theta \sim \pi$.

→ Ex: pour un modèle gaussien : $X_i \sim \mathcal{N}(\mu, \sigma^2)$, on pourrait par exemple choisir pour $\theta = (\mu, \sigma^2)$ la loi $\pi(\theta) = \pi(\mu) \times \pi(\sigma^2) \equiv \mathcal{N}(\mu_0, \tau) \times \text{InvGamma}(\alpha, \beta)$.

Définition : Les paramètres de la loi a priori (ici μ_0, τ, α et β) sont appelés des hyperparamètres. Ces hyperparamètres ont des valeurs fixées par le statisticien, on ne les estime pas !

Définition : Un modèle bayésien est la donnée, pour une v.a. (ou une suite de v.a.) d'une loi conditionnelle et d'une loi a priori :

$$X \sim f(X|\theta)$$

$$\theta \sim \pi$$

② - Loi jointe et loi a posteriori.

②

Etant donné que θ admet une loi (a priori), on peut exprimer la loi jointe pour un modèle statistique donné, à savoir

$$\pi(x, \theta) = f(x|\theta) \pi(\theta).$$

Rp: en pratique, la loi jointe n'a que peu d'intérêt en statistique bayésienne. On s'intéresse davantage à la loi a posteriori:..

Définition (loi a posteriori): On peut séparer les cas en 4 catégories:

1) La loi de X et la loi de θ sont discrètes: la loi a posteriori vaut

$$P(\theta = \theta_i | X = x) = \frac{P(X = x | \theta = \theta_i) P(\theta = \theta_i)}{P(X = x)} = \frac{P(X = x | \theta = \theta_i) P(\theta = \theta_i)}{\sum_k P(X = x | \theta = \theta_k) P(\theta = \theta_k)}$$

2) La loi de X est discrète et la loi de θ est continue. Alors on a:

$$\pi(\theta | X = x) = \frac{P(X = x | \theta) \pi(\theta)}{\int_{u \in \Theta} P(X = x | u) \pi(u) du}$$

3) La loi de X est continue et la loi de θ est discrète. On obtient

$$P(\theta = \theta_i | x) = \frac{f(x | \theta = \theta_i) P(\theta = \theta_i)}{\sum_k f(x | \theta = \theta_k) P(\theta = \theta_k)}$$

4) La loi de X est continue et la loi de θ aussi. Alors on a:

$$\pi(\theta | x) = \frac{f(x | \theta) \pi(\theta)}{\int_{u \in \Theta} f(x | u) \pi(u) du}$$

Rp: On résume souvent ces 4 cas par le cas continu/continu; c-à-d:

qu'on appelle loi a posteriori la loi dont la densité est donnée par

$$\pi(\theta|x) = \frac{f(x|\theta) \pi(\theta)}{\int_{\theta \in \Theta} f(x|\theta) \pi(\theta) d\theta}$$

→ Le dénominateur correspond à la loi marginale de X . On le note souvent $m(x)$. Ce dénominateur joue le rôle de constante de normalisation pour la loi a posteriori : en effet, il est indépendant de θ .
pour retrouver une densité.

→ Pour déterminer le comportement de la loi a posteriori, on travaille souvent à cette constante près, i.e.

$$\pi(\theta|x) = \frac{f(x|\theta) \pi(\theta)}{\int_{\theta \in \Theta} f(x|\theta) \pi(\theta) d\theta} \propto f(x|\theta) \pi(\theta)$$

fonction de θ // de θ .

→ Rq: • Si on cherche le maximum de la loi a posteriori, il est inutile de calculer la loi marginale.

- x est le vecteur des valeurs observées en pratique, permettant d'intégrer l'expérience et de mettre à jour la loi "a priori" qui devient "a posteriori".
- On donne le nom de statistique bayésienne car le lien avec le Théorème de Bayes est évident. (Exercice - Annexes Chap. 2 ex. 12, 14)

III - Plus loin sur la loi a priori

① - Comment la choisir ?

C'est souvent l'étape critique en statistique bayésienne.

Pourtant, c'est une étape fondamentale, et c'est même ce qui constitue la plus grande différence avec la statistique fréquentiste.

En général, ce choix peut être motivé par plusieurs types de considération:

- basé sur des expériences similaires du passé,
- basé sur un avis d'expert ou une intuition,
- basé sur la faisabilité des calculs (même si aujourd'hui les méthodes par simulation de type MCMC permettent de contourner cet écueil),
- basé sur la volonté de n'apporter aucune info nouvelle pouvant biaiser l'estimation.

② - Loi a priori conjuguée.

Définition : Une famille \mathcal{F} de lois de probabilité sur Θ est dite conjuguée pour le modèle statistique $\{f(x|\theta) : x \in \mathcal{X}, \theta \in \Theta\}$ si, pour tout $\pi \in \mathcal{F}$, la loi a posteriori $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ appartient aussi à \mathcal{F} .

→ Exemple: modèle Poisson-Gamma:
$$\begin{cases} X_i \sim \mathcal{P}(\theta) \\ \theta \sim \text{Gamma}(\alpha, 1) \end{cases}$$

⇒ Faire le calcul...

→ Intérêt principal: choisir une loi a priori conjuguée permet d'obtenir une loi a posteriori explicite. En pratique, la mise à jour se fait à travers les paramètres de la loi a priori, ce qui facilite aussi l'interprétation du changement dû à l'accumulation de l'expérience.

→ Remarque: Une loi conjuguée peut être déterminée en considérant la forme de la vraisemblance $f(x|\theta)$ et en prenant une loi a priori de la même forme (vue comme une fonction du paramètre).

→ Quelques exemples de lois conjuguées:

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
$\mathcal{NP}(\theta, \sigma^2)$	$\mathcal{NP}(\mu, \tau^2)$	$\mathcal{NP}\left(\frac{\sigma^2\mu + \tau^2}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$
$\mathcal{B}(\theta)$	$\text{Gamma}(\alpha, \lambda)$	$\text{Gamma}(\alpha+x, \lambda+1)$
$\text{Gamma}(r, \theta)$	$\text{Gamma}(\alpha, \lambda)$	$\text{Gamma}(\alpha+r, \lambda+\alpha)$
$\mathcal{B}(n, \theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha+x, \beta+n-x)$
$\mathcal{NP}\left(\mu, \frac{1}{\sigma}\right)$	$\text{Gamma}(\alpha, \lambda)$	$\text{Gamma}\left(d+\frac{1}{2}, \lambda + \frac{(\mu-n)^2}{2}\right)$
Pareto	Gamma	

→ L'impact de la loi a priori s'émouisse quand $n \rightarrow +\infty$. (illustrations ?)

③ - Loi a priori non-informative.

C'est le cas où l'on a peu (ou pas) d'information sur θ . On cherche donc à ce que les données parlent d'elles-mêmes au maximum, et de ne pas influencer la loi a posteriori. On étudie ici 2 familles de lois a priori non-informatives: la loi de Laplace et celle de Jeffreys. (le cas le + évident étant la loi uniforme...)

a) Loi de Laplace:

Définition: La loi a priori de Laplace consiste à poser $\pi(\theta) \propto \mathbb{1}_{\{\theta \in \Theta\}}$

C'est donc la loi uniforme (discrete ou continue selon Θ), voire la mesure de Lebesgue dans le cas d'une loi impropre.

→ Cette loi présente toutefois quelques défauts $\left\{ \begin{array}{l} \text{elle peut mener à une loi a priori impropre} \\ \text{elle est non-invariante par reparamétrisation} \end{array} \right.$

→ Ex: prenons le modèle Exponentiel (λ) , $\lambda > 0$; ainsi que sa reparamétrisation $\lambda = \exp(\theta)$, $\theta \in \mathbb{R}$.

- Dans le premier cas, on a $\pi_1(\lambda) \propto \mathbb{1}_{\{\lambda > 0\}}$

- Dans le 2^e cas, on a $\pi_2(\theta) \propto 1 \Rightarrow \pi_2(\lambda) \propto \frac{1}{\lambda} \pi_2(\ln(\lambda)) \mathbb{1}_{\{\lambda > 0\}} \propto \frac{1}{\lambda} \mathbb{1}_{\{\lambda > 0\}}$

⇒ 2^e cas fortement informatif: la notion de non-informatif dépend de la paramétrisation

b) A priori de Jeffreys:

(4)

On a vu qu'une bonne notion de loi a priori non-informative est une loi invariante par reparamétrisation. (ce qui n'est pas forcément le cas d'une loi Uniforme en fonction de la paramétrisation du problème).

L'a priori de Jeffreys est fondée sur l'information de Fisher.

Définition : La loi a priori de Jeffreys est donnée par

$$\pi(\theta) \propto |I(\theta)|^{1/2}, \text{ où } |A| \text{ est le déterminant de } A.$$

- Cette loi est invariante par reparamétrisation.
- Elle peut conduire à des lois a priori impropres.
- Elle est à éviter quand $\dim(\Theta) > 1$.
- $I(\theta)$ est un indicateur de la quantité d'information apportée par le modèle $f(x|\theta)$: il est donc intuitif que les valeurs de θ pour lesquelles $I(\theta)$ est plus grande doivent être plus probables a priori.
- Remarque sur la reparamétrisation: soit $\phi = h(\theta)$ avec h un C^1 -difféomorphisme. En notant π la loi a priori de θ , alors ϕ est de loi $\tilde{\pi}$ avec
$$\tilde{\pi}(\phi) = \pi(\phi) |(h^{-1})'(\phi)|$$
 - De plus on a $\tilde{I}(\phi) = I(\phi) |(h^{-1})'(\phi)|^2$, donc
$$\tilde{\pi}(\phi) \propto \sqrt{\tilde{I}(\phi)}$$
 (cela justifie la racine carrée).

→ Ex?

IV - Inférence bayésienne

Le résultat du travail bayésien donne accès à la loi a posteriori $\pi(\theta|x)$. Cette loi contient beaucoup d'information sur θ (beaucoup plus qu'en fréquentiste), et en pratique on résume cette loi à des indicateurs plus simples, comme :

- le maximum a posteriori (MAP),
- la médiane a posteriori,
- la moyenne a posteriori,
- un quantile a posteriori d'ordre p .

① - Estimateur bayésien

Définition : Si $d(\cdot, \cdot)$ est une fonction de perte, le risque bayésien de l'estimateur $\hat{\theta} = t(X)$ du paramètre θ pour la perte d est

$$R^B(\hat{\theta}) = E_{(\theta, X)}[d(\hat{\theta}, \theta)] = E_{(\theta, X)}[d(t(X), \theta)].$$

Rp: On peut généraliser cette définition à une fonction de θ , $\psi(\theta)$, avec un estimateur de cette fonction de θ .

N.B.: On intègre sous la loi jointe (θ, X) .

Proposition : Le risque bayésien est la moyenne du risque fréquentiste / fréquentiel suivant la loi a priori π . Ainsi,

$$R^B(\hat{\theta}) = E_{(\theta) \sim \pi}[\underbrace{R(\hat{\theta}, \theta)}_{\text{risque fréquentiste}}] = \int R(\hat{\theta}, \theta) \pi(\theta) d\theta.$$

Nous allons maintenant aborder le sujet d'un estimateur optimal au sens du risque bayésien.

(5)

Définition : Étant donnée une fonction de perte $d(\cdot, \cdot)$, l'estimation bayésienne associée à cette fonction de perte est définie, pour tout $x \in \mathcal{X}$, par

$$t^B(x) = \underset{t}{\operatorname{argmin}} \mathbb{E}[d(t, \theta) | X=x].$$

L'estimateur bayésien est donc $\hat{\theta}^B = t^B(x)$.

Théorème : L'estimateur bayésien est l'estimateur qui minimise le risque bayésien.

Proposition : L'estimateur bayésien associé au risque quadratique est donné par l'espérance a posteriori: $\hat{\theta}^B = t^B(x) = \mathbb{E}_{(\cdot|X)}[\theta | X]$.

Rq: On retrouve le fait que pour minimiser un critère de moindre carré, la meilleure estimation est la moyenne (cf régression linéaire).

- Si le risque est le risque L^1 (perte en valeur absolue), l'estimateur bayésien est donné par la médiane de la loi a posteriori $(\cdot | X)$.
- On peut généraliser tous ces résultats à une fonction de θ , en remplaçant θ par $\psi(\theta)$ partout. Exercices - Ardoise Prop 3. en 14, 15, 17, 18, 20

② - Intervalle de crédibilité.

La notion est proche de celle de l'intervalle de confiance, mais elle est \neq .

En statistique fréquentiste, on rappelle qu'un intervalle de confiance a des bornes aléatoires. Le niveau $(1-\alpha)\%$ de cet intervalle correspond à la proportion $(1-\alpha)\%$ que n réalisations de cet IC contiennent le vrai paramètre lorsque $n \rightarrow \infty$.

a) Région α -crédible:

Définition : Pour une loi a priori π donnée, un ensemble $C_\alpha \subset \Theta$ est un ensemble α -crédible si

$$P_\pi(\theta \in C_\alpha | x) \geq 1 - \alpha.$$

→ Interprétation: $P_\pi(\theta \in C_\alpha | x) = \int_{\{\theta \in C_\alpha\}} \underbrace{\pi(\theta | x)}_{\text{loi a posteriori}} d\theta$

On intègre donc l'expérience observée x dans la définition de cet intervalle.

b) Intervalle de crédibilité

On simplifie les choses en considérant que θ est scalaire.

Définition : Pour une loi a priori π donnée, un intervalle $IC_\alpha \subset \mathbb{R}$ est un intervalle de crédibilité de niveau $1 - \alpha$ si

$$P_\pi(\theta \in IC_\alpha | x) = 1 - \alpha.$$

→ Souvent on utilise des intervalles de crédibilité symétriques, c-à-d :

$$IC_\alpha(x) = \left[q_\pi\left(\frac{\alpha}{2}, x\right), q_\pi\left(1 - \frac{\alpha}{2}, x\right) \right], \text{ avec}$$

$$q_\pi(p, x) = \inf \left\{ u \in \mathbb{R} : P_\pi(\theta \leq u | x) \geq 1 - \alpha \right\}.$$

→ Comparaison entre intervalle de confiance | intervalle de crédibilité:

Exercice : Avenir Chap 2
22-13

- IC est aléatoire
- La prochaine réalisation de I sera avec $(1 - \alpha)\%$ de chance de contenir θ_0 .

- θ est aléatoire
- Ayant observé x , il y a $(1 - \alpha)\%$ de chance que IC_α contienne θ_0 .

③ - Loi prédictive a posteriori:

⑥

On dispose de $x = (x_1, \dots, x_n)$.

On souhaite prédire la prochaine observation x_{n+1} .

→ En fréquentiste, on a tendance à utiliser le prédicteur $E[X]$, avec $X \sim f(\cdot; \hat{\theta})$ et $\hat{\theta}$ estimateur de θ .

⇒ On ne tient donc pas compte de l'incertitude d'estimation sur θ .

Définition : On appelle loi prédictive a posteriori la loi de densité

$$\pi(x_{n+1} | x) = \int f(x_{n+1} | \theta) \pi(\theta | x) d\theta.$$

Comme en fréquentiste, on utilise ensuite l'espérance sur cette loi pour définir le prédicteur bayésien :

$$\hat{x}_{n+1} = \int x_{n+1} \pi(x_{n+1} | x) d_{x_{n+1}}.$$

II - Simulations

Enfin, pour conclure ce chapitre, il est essentiel de mentionner que des méthodes simulationnelles performantes (mais coûteuses en calcul) existent pour approcher la distribution a posteriori du paramètre θ (la loi $\pi(\cdot | x)$) dans un cadre général (pas forcément une loi a priori conjuguée, ...).

Ces techniques dépassent le cadre de ce cours, mais l'étudiant intéressé pourra notamment approfondir :

→ les techniques basées sur la simulation Monte Carlo

→ les algorithmes plus spécifiques de type MCMC :

- Algorithme de Metropolis-Hastings,

- Algorithme de Gibbs (échantillonneur)

→ les modèles bayésiens hiérarchiques.